LambdaStation:

Exploiting Advanced Networks in Data Intensive High Energy Physics Applications

Contact: Donald L. Petravick, Fermilab
petravick@fnal.gov.
Phone (630) 840-3935
Cell (630) 917-0728
fax(630) 840-6345

Focus Element: A. SciDAC Program

LambdaStation:

Exploiting Advanced Networks in Data Intensive High Energy Physics Applications

1 Abstract:

High Energy Physics collaborations consist of hundreds to thousands of physicists and are world-wide in scope. We propose to integrate the DOE Science UltraNet into High Energy Physics' large production systems in a general way allowing full-scale data systems to exploit dynamically provisioned optical links. These links are on a track to provide throughputs of 10 Gbps between our facilities and will allow us to aggregate our scientific network traffic onto congestion-free and almost loss-less connections.

To allow this exploitation of dynamic links, we will develop a forwarding and admission control service (a LambdaStation) for interfacing our production networks to research networks. Having the LambdaStation, we will introduce appropriate awareness of advanced networking into our storage production system services.

We will study and characterize application and network traces to allow us to model load and predict performance. We also expect that integration will influence the further development of production systems, which currently can not hope to exploit these links without substantial aggregation.

We begin with a particular focus on the experiments and facilities able to exploit ESnet and the DOE Science UltraNet research network: Fermilab, the joint CDF/CMS Tier 2 center at UCSD, and the CMS Tier-2 center at Caltech. We hope to include the US LHC Edge Computing at CERN, using the DOE-funded trans-atlantic link, and to generalize our solution to accommodate other dynamically provisioned links, and other communities.

LambdaStation:

Exploiting Advanced Networks in Data Intensive High Energy Physics Applications

Scientific Discovery through Advanced Computation
Solicitation DE-FG01-04ER04-02, LAB 04-03

2 Participants

DOE Laboratory Contact:
Don Petravick, Fermilab, MS 370, P.O. Box 500, Batavia, Illinois 60510, Phone (630) 840-3935, Cell (630) 917-0728, fax(630) 840-6345

List of Participants

California Institute of Technology:
- Harvey Newman (CO-PI)
- Julian Bunn

University of California at San Diego:
- Frank Wuerthwein (CO-PI)
- James G. Branson

Fermi National Accelerator Laboratory:
- Don Petravick(PI),
- Phil Demar,
- Lothar Bauerdick

Northwestern University:
- Peter Dinda

3 Background and Significance:

3.1 Background:

Particle Physics Collaborations consist of hundreds to thousands of physicists and are world-wide in scope. The SciDAC Particle Physics Data Grid Collaboratory Pilot (PPDG) project develops, acquires and delivers vitally needed Grid-enabled tools for data-intensive requirements of these experiments. To fully exploit the science potential latent in their data, CDF and D0 at Fermilab and BaBar at SLAC are expanding their data analysis to integrated distributed systems based on Grids. Moreover, U.S. physicists preparing for the analysis of data from the CMS and Atlas detectors at the Large Hadron Collider at CERN (LHC) face unprecedented challenges: (1) massive, globally distributed datasets growing to the 100 petabyte level by 2010; (2) petaflops of distributed computing; (3) collaborative data analysis by global communities of thousands of scientists. PPDG, together with the NSF-funded iVDGL and GriPhyN projects, is moving to the development of next generation of integrated Grid systems to meet these challenges, and to fully exploit the LHC's potential for physics discoveries. Today, all these high energy physics PPDG experiments' grid systems are limited by their treatment of the network as an external, passive, and largely unmanaged resource. Moreover, to date, no advanced network linking the U.S. HENP Laboratories and key universities involved in Grid and network development has been available to research and prototype solutions to these limitations.

Another important use for very high throughput networks is to move the LHC data across the Atlantic from CERN in Geneva, Switzerland, to the U.S. Tier-1 regional centers, Fermilab for the CMS experiment and Brookhaven for Atlas. From there data will be distributed to Tier-2 regional centers at Universities like Caltech and UCSD. These data transfer facilities will have components of a quasi-real-time system as data taken at the LHC will have to be continuously distributed to the regional centers. Data streams of raw detector data and reconstructed data ready for physics analysis are being spread over the distributed regional centers, selected and targeted to specific physics interests, to ensure full data access for U.S. physicists to LHC data and to serve analysis hotspots making data available to specific regional centers.

To ensure full connectivity of the U.S. to CERN and full access of U.S. scientists to LHC data, the U.S. LHC software and computing efforts have started to put up U.S. LHC Edge Computing elements at CERN with sufficient data caching and data selection resources and a 10Gbit connectivity from these systems across the Atlantic to the DOE funded link to CERN in Chicago. At both endpoints clusters of CPUs and storage elements are being used that are similar to the systems described above. LHC data taking will start in 2007, and the LHC experiments are conducting a program of work to scale up to the required throughputs and functionalities that employs yearly "data challenges" that to exercise the emerging end-to-end data flow systems to increasing degrees of complexity and size, starting with a 5%-sized data challenge in 2004.

3.2 Problem Area:

Over the past several years, there has been a great deal of research effort and funding put into the deployment of optical-based, advanced research networks, such as National Lambda Rail, DataTag, CAnet4, Netherlight, UKLight, and most recently, the DOE Science UltraNet. These advanced

network infrastructures potentially have the capacity and capability to meet the extremely large data movement requirements of the Particle Physics Collaborations. To date, the focus of research efforts in the advanced network area have been primarily to provision, dynamically configure & control, and monitor the wide area optical network infrastructure itself. Application use of these facilities has been largely limited to demonstrations using test stands or small prototype high performance computing systems. The issue of integrating existing production-use computing facilities connected to the local network infrastructure with advanced, high bandwidth research networks has remained unaddressed. Fundamentally, there is a last mile problem between HEP experiment production-use computing facilities and the advanced research network infrastructure. Our proposal addresses that void.

3.3 Proposed Solution:

We propose to enable our very large, production-use mass storage systems, running full-scale SciDAC applications to exploit advanced research networks. We will implement the necessary innovations in our local network and application environments to enable our SciDAC applications to send traffic, on a per-flow basis, across advanced network paths, specifically the DOE Science UltraNet. We will create the capability to selectively forward designated data flows between our capacious storage systems across our local network infrastructure and a dynamically provisioned UltraNet path. Concurrently, other traffic flows from the same storage system would be forwarded across the same local network infrastructure onto our conventional, routed wide area network path. This will allow us to use the multi-million dollar installed base of oft-renewed computing facilities running the applications we will study and develop. Our project is intended to incur no additional equipment cost to the storage systems and very small additional equipment costs to enhance the local network infrastructure.

The ability to selectively forward traffic on a per-flow basis requires us to develop the capability to dynamically reconfigure forwarding of specific flows within our local production-use routers. Investigating how to do this, we have determined that suitable infrastructure to accomplish this is missing. We will develop that infrastructure. We refer to it as LambdaStation. If one envisions the optical network paths provided by UltraNet, National Lambda Rail, and other advanced optical-based research networks as high bandwidth data railways, then LambdaStation would functionally be the railroad terminal that regulates which flows at the local site get directed onto the high bandwidth data railways.

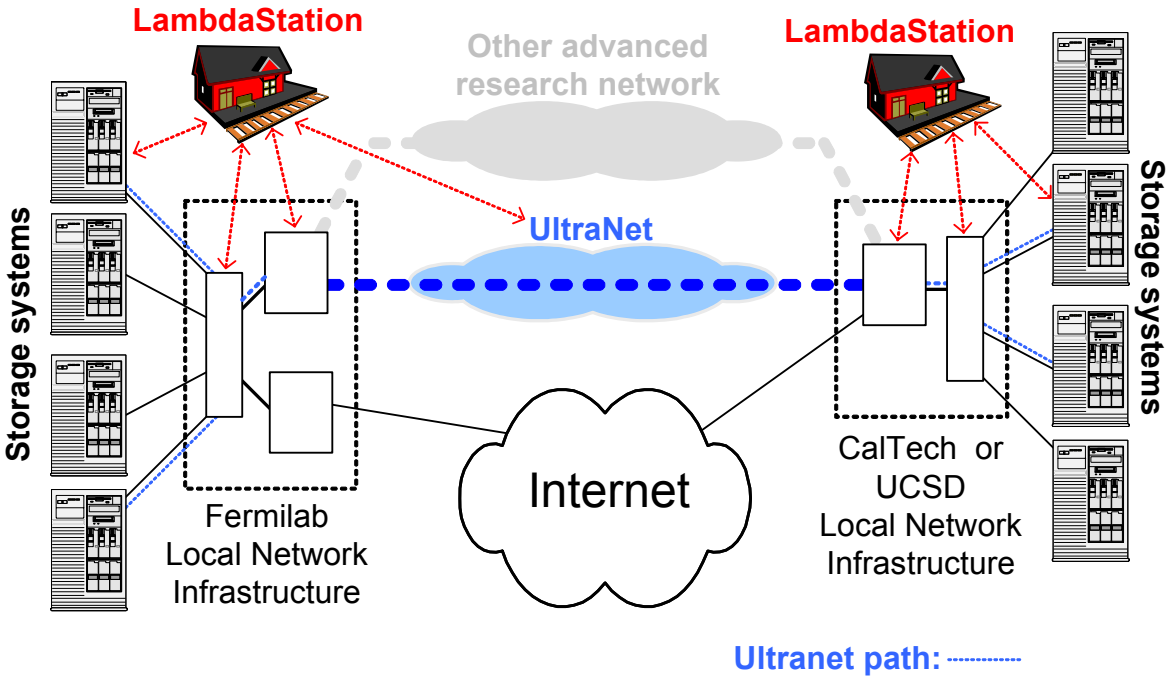

Figure 1: LambdaStation

LambdaStation would coordinate with UltraNet on network path availability, scheduling, and setup, direct appropriate forwarding within the local network infrastructure, and provide the application with the necessary information to utilize the high bandwidth UltraNet path. Having created LambdaStation, we will introduce awareness and exploitation of advanced networking into data management services of our experiments. Since our focus is on the experiments and facilities able to exploit the DOE Science UltraNet research network, our proposal is centered on Fermilab facilities, and the joint CDF/CMS Tier 2 center facilities at Caltech/UCSD. We will work to include the DOE funded Edge Computing for US CMS at CERN. These capabilities will, moreover, naturally transition to be available to users of the Fermilab central storage facilities.

3.4 Significance of Research:

Our proposal covers three areas that must be successfully addressed before optical-based advanced network technologies can be of significant benefit to our Particle Physics Collaborations, or even within the broader scientific research community in general:

- **Dealing with the local network last-mile problem.** Our proposal seeks to adapt existing, production-use local network facilities to support access to advanced research networks. Until this problem is properly addressed, access to advanced research networks will be limited to demonstration computing systems and facilities with customized access. That type of approach does not scale well in distributed computing environments based on large numbers of commodity computing platforms.
- **Bringing real-world, production-use facilities and applications into the advanced research network environment.** Our proposal promotes the necessary synergy between large scale computing facilities and the capabilities of advanced network technologies to advance the use of the latter by the former.

- **Developing flow-based alternate network path selection capabilities.** Our proposal represents a preliminary step toward special or alternate forwarding of specific flows for performance or policy reasons. Assuming the long term vision of collaborative wide area networking is one of many lambdas coupled with policy constraints, flow-based path selection is a highly desirable capability to investigate.

3.5 Relevance to Office of Science Research Programs:

Our proposal advances the research efforts of the Office of Science by:

- Bringing the advanced research network capabilities of the Office of Science UltraNet Project to HEP experiments that are deploying applications and middleware in a Grid and DOE Facility environment.
- Bringing multi-million dollar installed base of oft-renewed mass storage computing facilities into use for Office of Science advanced network research and development. Our project enables the Office of Science's research efforts to increase the benefit from existing, high performance computing facilities.

3.6 Near Term Significance; An End-to-end Use Case:

How do we imagine LambdaStation integrating into a physicist's work environment? It is a mechanism enabling middleware to use dynamically provisioned optical links. Part of our research is to develop the knowledge of how to specify LambdaStation's interaction with storage systems, grids, and networks. The near term impact of our proposal is best described by what we imagine a typical use case to be. We illustrate this with a non-trivial motivating example that is within our collaboration's means to implement.

Consider a demonstration of updating files for CMS on an interactive analysis cluster at the UCSD Tier-2 center from a master copy at the Fermilab Tier-1 center. This might typically consist of 3.5 TB of data, spread over a storage system of 50 file servers on each side. The data is normally packed in one gigabyte files, so 3500 files would be transferred. Properly aggregated, the entire file transfer could be completed in slightly less than an hour across a 10Gbps UltraNet path.

The SRM (Storage Resource manager) residing in the storage system at FNAL is presented with the list of the 3500 files to transfer. Using knowledge of the available bandwidth between the sites, the type of network, the bandwidth of the disks, and the loading of the storage system from other uses, file transfers begin. Initially this will use the pre-existing conventionally routed network.

UltraNet and LambdaStation then announce a dynamically provisioned, full-duplex, layer-2 connection between Fermilab and UCSD. Initially, the coordination of the transfer flows will be done manually. Automatic coordination of this is future work, which relies on the development of automatic provisioning in UltraNet.

Given the announcement of the provisioning, the two SRM's adjust the file-wise parallelism of the 3500 file transfer. Since, unlike routed networks, the dynamic link is congestion–free and ought be practically loss-less other aspects of the file transfer might change. For example, the underlying file transfer protocol used may change.

The storage system notifies LambdaStation of the network-level flows associated with the file transfers. The notification includes sufficient information for configuring the local network to forward appropriately, as well as control admission to and from the dynamic link. The notification also has other attributes, such as the virtual organization on whose behalf the file transfer is taking place. LambdaStation returns a positive acknowledgement for appropriate flows.

LambdaStation software at UCSD and Fermilab configures local switches and routers to forward outbound flows onto the dynamic link. LambdaStation software configures routers to admit appropriate inbound flows, and reject inappropriate ones.

Another Virtual Organization(VO), CDF, which is also active at FNAL and UCSD, happens to be transferring files to UCSD at the same time as CMS. Since the UltraNet link is configured for the exclusive use of the CMS VO, the LambdaStation must not forward CDF flows onto the dynamic link, and the corresponding SRM's ought not receive a positive acknowledgement to these notifications.

Because we propose to integrate UltraNet to last-mile production networks during demonstrations, LambdaStation also monitors and asserts that the expected flow behavior is occurring and protects the production network from certain errors. An example error might be accidental reconfiguration of UltraNet, with flows from PNNL to Oak Ridge being forwarded to the FNAL network during our one hour demonstration. The LambdaStation would not allow these flows to enter the FNAL network.

At the end of an hour, very many, but not quite all, files have transferred. The dynamic link is scheduled to be torn down. The local LambdaStations reconfigure local forwarding back to the production network path. The SRM's become aware of the reduced bandwidth; and use different parallelism and possibly different protocols for any remaining file transfers.

3.7 Long Term Significance:

We anticipate success and assume a growing role for dynamically provisioned optical links as alternate paths for our high bandwidth applications. Our proposal is based on a long term perspective that requires a design flexible enough to:

- Operate in a world where optical paths are relatively scarce today, but increasingly available in the future.
- Interoperate with advanced research networks other than DOE Science UltraNet.
- Interoperate with facilities having local network infrastructure that is different than ours.
- Interoperate with a great number of facilities and a great number of end-node at each facility.
- Benefit from any relevant resource prediction models we can discover.
- Interoperate with a heterogeneous mix of Virtual Organizations (VOs).
- Converge to standards as they emerge.
- Integrate with GRID middleware as it evolves.

4 Preliminary Studies:

In preparation of this proposal we searched for relevant work, and re-surveyed the systems used at CDF and the US CMS prototypes that can benefit most from an integration with UltraNet.

4.1 Relevant Work

We have followed the development of the GGF informational document which relates optical networks, dynamic path provisioning and grid software and systems [GGF1]

We have visited iCAIR [ICAIR] and are aware of ongoing work there, including the Optical Dynamic Intelligent Network (ODIN) service layer, THOR TeraScale High Performance Optical Resource-Regulator , and DEITI Dynamic Ethernet Intelligent Transit Interface provides for L2 extensions across Ethernet links at the network edge. We understand these projects to be layer-2, not layer three concepts. [PHOTONIC]. We have reviewed EMERGE[EMERGE] with iCAIR, and found that it deals with quality of service, not path selection

We understand GMPLS to be an important industry direction, but need to investigate how immediate its application might be for practical integrations, such as ours, and especially for our initial integrations which will multiplex many modest flows onto a 10 Gigabit link [GMPLS1-3]

We have visited with Cees De Laat, and discussed of work done in the AAA [AAA1][AAA2] framework for controlling switching , in a context with authentication, authorization and policy requirements similar to our own [AAAQoS], and understand that it illustrates the use of potential software building blocks, but does not constitute a solution to our problem.

We are aware of GARA, a component of the Globus [Globus] framework dealing with resource allocation, and its application to quality of service.

We have participated in the DOE ESnet workshops[ESNET], and have hosted a DOE Science UltraNet meeting at Fermilab [FNALMEET]. UltraNet will provided an advanced, wide area network and test bed for our research. The DOE UltraNet will provide a control plane using GMPLS and TL1.[ULTRANET}

We have tested policy-based routing within our local network infrastructure, and successfully forwarded select flows across an alternate wide area path, using either DSCP code points or source/destination address/port tuples to differentiate the flows.

We have a storage system, dCache [DCACHE], which Fermilab has developed jointly with the DESY laboratory in Hamburg, Germany. The system supports a SciDAC Storage Resource Manager (SRM) project interface which we have developed jointly with LBNL, amongst others. The co-Principal Investigator for SRM is on this proposal[SRM]. The system is a major component for the CDF experiment at Fermilab [CDF] and the CMS experiment at the LHC, both of which are participating in the PPDG SciDAC project. CMS is also participating in the LHC Computing Grid project [LCG]. The SRM and dCache-based Grid storage element has been developed together with U.S. CMS. It has been tested and deployed within the U.S. Grid2003 environment

[Grid2003] and has been included as a component in the LCG system. The system has a GridFTP interface, and we have worked on protocol extensions for GridFTP [GRIDFTPV2].

To achieve consistent high performance, an application must either adapt its behavior to changing resource availability, or make resource reservations. We have investigated the tradeoff between resource reservations and application adaptation Many adaptation mechanisms have been proposed, such as path selection, Replica selection [Myers1], Source routing, Multipath routing, and Configurable overlay topologies [Chu1, Sundararaj1].

The control systems that can make use of such mechanisms in pursuit of application goals must be able to estimate and understand the time costs of different options and their tradeoffs. These estimates or predictions depend not only on measures of the available network resources such as host availability and load [Dinda2] and available network bandwidth [Dinda3, Qiao1, Wolski1], but also on the actual application demands, such as the length of jobs [Dinda1, Harchol1] and the size of flows [Zhang1, Guo1]. Unfortunately, it is often difficult or impossible for application developers to supply this information. For example, the amount of data produced by a query in a data-intensive high energy physics application depends on the selectivity of the query, which is a property of the domain of physics, not of computer science.

A similar situation exists in situations where adaptation is augmented or replaced with resource reservations. To make use of a reservation mechanism, the application must supply a profile of its planned use and the expected performance. The network uses this information to conduct admission control and to throttle the application to its supplied profile. Reservation mechanisms, both for networks and hosts, are well understood but determining the demands of the application remains a key difficulty shared with adaptation.

4.2 A Synopsis of High Energy Physics Production Computing.

At the hardware level, the overall technical direction in the HEP storage system field has been to construct a high performance distributed storage system by exploiting very large number of commodity PCs, or workstations, equipped with low cost disks. For CMS and CDF this trend is evident in that large amounts of disk are assembled from commodity computing. Storage system software is provided to make these disks appear as a coherent disk cache system. At Fermilab, the largest such storage system consists of 64 dedicated cache machines, representing 150 TB total disk capacity, moving over 30 TB of day in production use. Design attention is driven by both cost and scalability. Preference is given to designs that can serve as reference models for replicating similar systems at collaborating universities and other institutions. Cost-performance is given great weight, and the cost-benefit of the embedded network is considered along with the cost of storage and machines.

An exemplar of the kind of use that needs to be supported is an analysis cluster. Such a cluster is supported by a companion storage system. Current practice is to package physics data sets as an ensemble of a great number of relatively small files. High throughput is obtained by dividing the analysis up file-wise, with concurrent and independent access to a large number of files. An additional property of these systems is that the storage system is in constant use by its local cluster. Within the local cluster, access is via POSIX reads and writes over an IP network. External access moves whole files to and from the storage system over the wide area network. It is this staging activity that would exploit a dynamically provisioned optical network. For the analysis cluster use

case, we begin our research with reference local system designs where wide-area staging and local POSIX access compete for disk bandwidth. In this model, any given file transfer has no hope of proceeding at 10 Gbps, though these clusters can easily sink such a rate in aggregate.[CDF][LCG]

Such is the state of current HEP storage systems. Availability of very high bandwidth optical-based wide area networks, evolution and progress by the networking and storage industries, and our own work might eventually result in different thinking about system implementations and design. But, recent history would indicate this model will be dominant for the next several years. We begin our research integration of experiment distributed systems and optical networks in an environment that includes the following salient features:

- Storage system hardware and access patterns having very large ensemble throughput, but only modest host transfer files rates, compared to 10Gbps. Cost effectiveness dictates that individual machines have network interfaces of modest speed compared to 10 Gbps
- Our Facilities will be directly interconnected at layer-2 by optical networks, but reaching storage end systems will require transiting the local site, production-use, routed network infrastructure. Multi-homing storage end systems to provide a physically separate network path into research networks does not scale well, both in terms of the number of end systems in use and the potential for connecting to multiple research networks, each requiring its own storage end system NIC.
- We have source-code control of intelligent IP-based storage systems; including the ability to add new features. The storage system software is modular with respect to layer-5 transfer protocols, and has been used to test layer-4 innovations, such as FAST [NetL1].
- We have reasonable access to very large systems running full-scale, evolving, SciDAC applications with which to exploit research networks. Use of these storage resources is constrained only by a requirement that concurrent access to production networks be maintained and that hooks to the research network not be disruptive to production use of these systems.
- We have access to the U.S. LHC Edge Computing systems at CERN that are connected to the DOE transatlantic link to CERN in Chicago, and that are available for specific tests and simulations of LHC-scale data movement and access

5 Research Design and Methods:

We have four research objectives:

1. Application Integration:- Understand and demonstrate how to optimize the experiments' grid-based data analysis systems by modifying selected services, including large scale storage services, to exploit advanced research network paths. The emphasis will be placed on end-to-end, high performance data transfers. To the extent supported by the UltraNet, our systems will manage data flows and select from dynamically provisioned research network paths. We will also apply this to the US LHC experiments, specifically for the data transfers from the U.S. LHC Edge Computing systems at CERN to the U.S. LHC Tier-1 centers, and between the U.S. LHC Tier-1 and Tier-2 centers, and to the emerging Grid-enabled Analysis Environment [GAE] being developed at Caltech as part of the PPDG project

2. Facility Integration:- Understand and demonstrate how to dynamically modify elements of existing local network facilities to connect production data storage systems to advanced research network(s) while concurrently retaining connectivity to existing production wide area network(s). The goal is to enable exploitation of research network path(s) on a per-flow basis.
3. UltraNet Integration:- Establish physical connectivity between the project sites and UltraNet at 10Gbps. Interact with UltraNet-specified scheduling and provisioning interfaces to establish and tear dynamically provisioned optical network paths between our respective sites.
4. Performance Study:- Study and characterize the application layer performance analysis of end-to-end flows, as well as the overall aggregate behavior across the research network path, when under load from these applications. We will characterize and predict the network traffic using end-to-end monitoring, tracking and analysis techniques, with the goals of helping applications to adapt to changing network conditions, and making effective decisions in directing larger traffic flows in real time, along appropriate (dedicated-lambda or shared) paths.

The project goal is to enable CDF and CMS applications to be aware of and able to exploit dynamically provisioned optical network paths offered by advanced networks. Our technical approach to accomplishing this goal has two distinct components. First, we will modify the storage system(s) in use, including enhancing the interfaces of these systems to the applications. The storage system currently in use is dCache [DCACHE], which includes GridFTP and the SciDAC funded Storage Resource Manager (SRM) interfaces [SRM]. DCache is accepted as an LHC Computing Grid (LCG) product, and has been used to test FAST. We have the capability to augment the system with layer-4 (advanced transport stack) and layer-5 (strategic end-to-end path building) protocols. We will also work in the context of the emerging Grid-enabled Analysis Environment being developed in PPDG and the US LHC experiments.

Second, we will adapt the existing local network infrastructure to support dynamic modification of forwarding path for select traffic flows to enable them to traverse advanced wide area networks. To accomplish this, we will develop a local path selection service, called LambdaStation, that will interface with the local storage system applications, coordinate the dynamic establishment of paths across UltraNet and other collaborating advanced wide area networks, and finally modify local network configurations to facilitate the appropriate path forwarding between storage system(s) and advanced wide area network.

5.1 Application Integration

We propose to begin by making the storage systems aware of the path selection service, since that enables the currently deployed systems of CDF and CMS experiments to utilize UltraNet. We will liaise with the general integration efforts of the experiments by working with them directly and via the SciDAC PPDG project.

The LHC-era storage systems that are being deployed have their own internal notions of scheduling and admission control. Experience has shown that these features are needed to prevent network congestion and disk-thrashing.

The SciDAC storage resource manager (SRM) exposes this to storage system applications through a web-services based management interface to storage systems. The dCache software currently used to manage storage in US CMS and CDF has this interface and functionality.

The SRM interface expresses queuing and load balancing to file transfer users by the pacing of delivery of Transfer URL's. SRM based systems also use internal queuing and load balancing features when performing storage-system to storage-system transfers, which can be requested and monitored though other aspects of the SRM web services API. We expect to investigate how these properties interact with the drastically varying bandwidth made available by dynamically provisioned links. We believe that projects like the SRM can benefit from the experience we will gain with this integration. Therefore, we must pragmatically bring these storage systems to the point they can exploit high performance networks in a stand-alone way. Areas of investigation include:

- How to provide for a hint allowing the transfers to be batched pending availability of a link.
- How to provide the ability to increase or decrease the number of transfers given the presence or absence of an alternate path link, or change in the status of bandwidth availability on an existing alternate path link.
- How to provide the ability to select layer 5 file transfer protocols (or tweak the parameters in layer-5 protocols) based on the presence or absence of an alternate path, or the performance characteristics of an alternate path link being used to reach a peer storage system.
- Understanding if these new capabilities are adequately expressed in the GLUE information schema used for integration with PPDG experiments.
- How to integrate with RPS or other systems providing a real-time model.

DCache is modular with respect to layer-5 file transfer interfaces. It currently supports FTP, kerberized FTP, GridFTP V1.0, and HTTP interfaces. At the flow level, dCache deployments are very much like an independent ensemble of linux computers, each with its own disk. We expect to be able to generate traffic loads that are not sustainable across our current production wide area network paths. We will likely be able to aggregate sufficient data flows to fill a 10 Gigabit pipe.

We have the capability to support a non-TCP layer 5 protocol, such as UDT. Our primary goal in this proposal, however, is to utilize the standard layer-5 protocols currently supported within the CMS and CDF production storage systems. Introducing UDT and similar research protocols into our storage systems is a research option open to investigation, but is not incorporated into this proposal's body of work.

5.2 Facility Integration Work

We will modify our local network infrastructure to bring the capacity and capabilities of advanced research networks into our existing local mass storage facilities, where the applications reside. Our goal is to selectively forward designated data flows from our capacious storage systems onto a dynamically provisioned UltraNet path, leaving our remaining traffic flows forwarded via our conventional wide area network path, ESnet.

Our analysis indicates the enabling facility extension we require is a local path selection service that can dynamically modify traversal paths within our production local network(s) to selective forward flows across dynamically provisioned links provided by UltraNet. While classical Internet thinking indicates that path selection based on destination and possibly source host address will suffice, we see several factors which motivate per-flow (i.e. protocol, source, source-port, destination, destination-port) forwarding decisions. Examples:

- We want our system to function on multiple, independent paths, and may have to balance admission by flow to balance the load.
- We may have to support the notion of ownership of a circuit by a VO, and constrain transfers on that circuit to the VO; however we can see deployments where data transfers between two IP nodes may be performed on the behalf of different VO's.

We also will investigate the notion that path selection may change for a flow during its lifetime. One motivation would be straggler flows from an aggregation that need to continue after a dynamic optical circuit needs to be torn down.

Our investigations have not yielded an extant path selection system with the requisite properties. Since this service is a cornerstone of our concept, we propose to build a system at least sufficient for our needs. We refer to it as LambdaStation. We understand its potential for general relevance, and so will attempt to use or interface to existing research (e.g. AAA) in this area of research. We also propose to continue the search for relevant work

The functional elements of LambdaStation are:

- Determine or maintain awareness of the availability of layer-2 paths within UltraNet. As an enhanced goal, the service design will be flexible enough to determine or maintain a similar awareness within other research networks relevant to the HEP SciDAC experiments.
- Coordinate the establishment and teardown of dynamic optical paths across UltraNet or other networks relevant to the HEP experiments. This might be a scaffolding feature that is useful while GRID middleware evolves to the point that dynamic circuit establishment & teardown can be done in context overall GRID resource allocation.
- Maintain awareness of application flows needing admission to UltraNet, or other advanced networks, potentially supported by a declarative interface integrated into middleware.
- Maintain awareness of the experiment's Virtual Organization system, for authentication and authorization relevant to admission decisions.
- Provide controls to enforce admission decisions and protect from intentional or accidental misuse.
- Depending on implementation, cooperate with application software on computers originating flows that could generate markers such as DSCP code points, port number assignment, or VLAN tags that can be used to differentiate per-flow router forwarding decisions.
- Perform appropriate configuration modification on local routers, using conventional policy routing mechanisms, to enable per-flow forwarding onto UltraNet or other alternate networks.

- Allocate resources, as a component of the broker function, in terms of managing appropriate levels of flows in an aggregation environment allowed across an alternate network path.
- Monitor local routers or other network equipment to assert that the expected flow patterns are occurring, and to deal with exceptions.
- Coordinate with remote LambdaStations for the establishment of symmetric end-to-end traversal paths across alternate paths.
- Investigate the relation of the LambdaStation to monitoring packages such as MonaLisa [MonA1], and the equivalent European monitoring infrastructure used by the LCG.

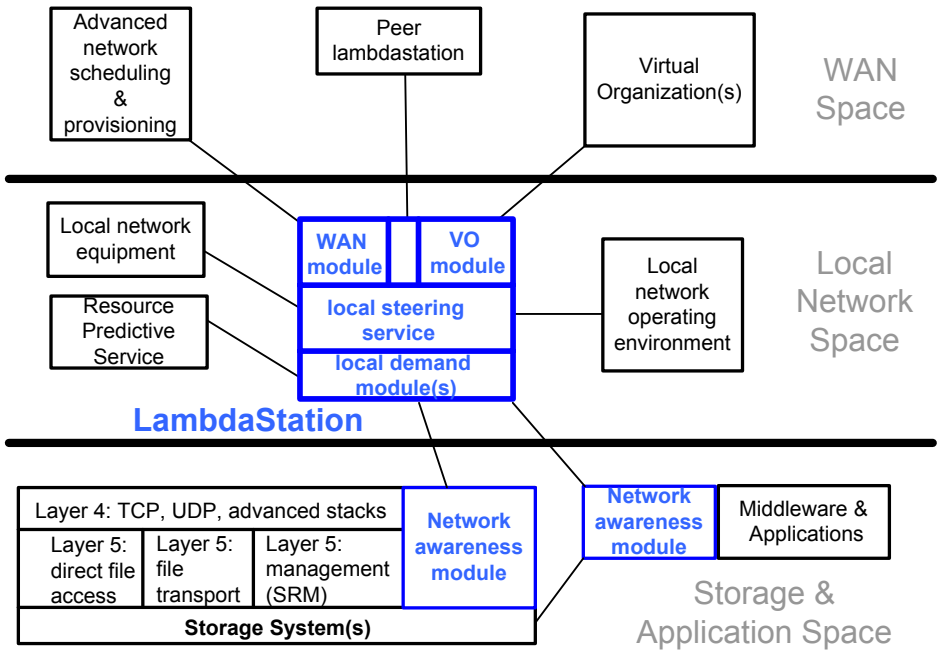

Figure 2 – LambdaStation Relationships

The initial deployment of our LambdaStation is based on having a production wide area network infrastructure augmented by one or more non-production advanced network paths available for select traffic. However, the model for this proposal is more generic. It is one of a default network path for normal production traffic, and alternate production network paths (potentially many in number) available for applications that are authorized to use them. We will investigate how to apply awareness of all the flows to the security aspects of our facility.

5.3 UltraNet Integration Work

We will bring our capacious test and production facilities, with their as-currently developed and evolving, applications to UltraNet. Caltech and UCSD will establish 10Gbps layer-2 paths to UltraNet at Sunnyvale. Fermilab will establish a 10Gbps layer-2 path to UltraNet at Starlight. Aggregated 1 Gbps links into UltraNet may be deployed on an interim basis, if the initial UltraNet service offering is less than a 10 Gbps backbone service.

We will liaise with the UltraNet engineering group. The exact program of work depends on UltraNet's implementation, including the nature of the interfaces where Fermilab and

Caltech/UCSD will meet UltraNet, and the Ultranet's schedule for provisioning bandwidth Areas of liaison with the UltraNet team will include:

- Investigating a transcontinental loop-back service from Fermilab to Sunnyvale and back. The loop-back service would be useful in emulating path latency found on trans-Atlantic links. The especially capacious and extensive computation, storage and network systems at Fermilab could be used as source and destination systems for the prototyping of very high bandwidth trans-oceanic data movement.
- Utilizing UltraNet as one component of a multi-network, dynamically provisioned optical path. For example, creating an optical path between project sites and CERN.
- Use of GMPLS, and other modes of bandwidth management, in developing the means to optimize end-to-end performance among a set of storage systems, and perhaps real-time processes, and other traffic flows.
- Have UltraNet provide appropriate monitoring capabilities that can be utilized to assist our facilities in achieving high performance across the network. Both individual application demand and the aggregate behavior of the network will be monitored.

Our proposal, while structured to make use of UltraNet, will be capable of utilizing any advanced research network that can provide a scheduling interface and an associated provisioning service. We are particularly desirous of reaching the Edge Computing at CERN.

5.4 Performance Study

We will investigate packet loss on our end-to-end flows, to ascertain how well dynamically provisioned optical links eliminate congestion as an issue for production data transfers where the end network is a well-engineered layer-3 path, and the intermediate network is a layer-2 optical circuit. Our results here will guide investigations needed to improve performance, including:

- Robust diagnosis of congestion and loss in such network paths.
- Investigation of alternate TCP stacks, building on the work begun by US CMS on FAST.
- Investigating a non TCP file transfer protocol, such as UDT.

We will analyze the performance and other salient features of our ensemble of high performance, routed and switched networks; very capacious compute and storage facilities; and high performance middleware and SciDAC applications.

As part of our study, we propose to develop and implement techniques that will predict the resource demands of applications as they run. If successful, these techniques will greatly expand the instances in which adaptation and reservation can be used, leading to consistently high delivered performance with high utilization.

We will use physics applications running on UltraNet as our driving applications. From instrumented applications we will collect traces with which we will assess the effectiveness of different predictive approaches. The best approaches will then be implemented to operate at run-

time. More specifically, we will log the timestamp, duration, and size of each TCP or UDP flow sourced or sinked by the application, as well as a time series, based on aggregated packet sizes, for each flow to capture time dynamics. In addition, we will collect application-level context, specifically the parameters of the highest level routines, flattened into time-stamped vectors. Additional information will be captured if needed, but we seek to work at the lowest semantic level possible to provide maximal generalization of our work

At this point, we contemplate exploring the following modeling approaches: traditional arrival process models, time series models, Markov models, and locally-weighted memory-based learning. Once we start examining the data, we may decide to emphasize particular approaches from the following or consider other approaches.

Modern results about arrival processes (early '90s to today) have demonstrated that the traditional arrival model is inadequate in virtually every domain of computer systems and networks. It has largely been supplanted by models where at least the sizes are drawn from power-law i.i.d distributions such as Paretos [Harchol1, Willinger1]. Power-law distributions have "memory", the exploitation of which has been profoundly changing the performance analysis community, causing reconsideration of questions as diverse as "starvation" in shortest remaining processing time scheduling and the value of process migration or restart.

Our work will characterize application flows on UltraNet within the traditional arrival process model, deriving distributions for both size and interarrival time. We will also test whether i.i.d. is a reasonable assumption here, and develop a model that includes serial correlation, if needed. This level of modeling is appropriate for admission control in resource reservation as well as global scheduling using queuing-theoretic analysis.

For a single flow, the rate may vary over time. For example, a high selectivity query might generate bursts of data. We will explore the use of time series analysis to characterize and predict the rate over time of individual and aggregated flows. This work will leverage the RPS system (http://rps.cs.northwestern.edu) and the Northwestern group's years of experience in statistical signal processing. Being able to predict flow rate as a function of time would useful in reservations, but also in adaptation.

We suspect that very high quality predictions of network demand will require that we derive an explicit model of the application, one that captures the bare essentials of its stages, and operations with respect to the network. We plan to explore the use of hidden Markov models (HMMs) for this purpose. In an HMM, the number of states and the transition probabilities are unknown, but can be estimated from observations. HMMs have been shown to perform amazingly well in predicting which file is likely to be accessed next even under constrained state space sizes represented as tries [Kroeger1]. We plan to apply a similar methodology to predict the next state of the application, and from that state, the likelihood of flows of different sizes being initiated.

In locally weighted memory-based learning, instances of function execution, maps from particular inputs to their corresponding outputs, are stored in a database. A query to predict function output from a given input is answered by selecting a "neighborhood" of instances "around" the query input and returning a weighted sum (or other kernel) of their outputs. In the context of the Purdue

PUNCH project, this approach was used very successfully to predict the CPU demands of submitted jobs based on their parameters [Kapadia1]. We plan to apply a similar approach to try to predict flow sizes and probability of flow initiation based on the current application state.

Approaches shown to be successful based on offline work with our trace data will be implemented as parts of the publicly distributed RPS system.

5.5 Management Structure

Our proposal provides a strategic direction for enabling data intensive high energy physics applications to exploit UltraNet, and other advanced research networks. Organizationally, our project management structure will have two components. The Steering Group will be responsible for the tactical decisions and directions necessary to implement our proposal's strategy. The Steering Group membership will be made up of PIs, and Co-PIs affiliated with the three Project sites, the Project Collaborations (CMS & CDF), and Northwestern University, the performance study arm of the Project. The PI for the Project, Don Petravick, will chair the Steering Group. Other Steering Group members will be: Lothar Bauerdick, Harvey Newman, Frank Wuerthwein, and Peter Dinda.

The Technical Group will be responsible for turning the Steering Group's decisions and directions into implementations, either via Group members directly, or staff working under them. The Technical Group will use project-funded effort. The membership of the Technical Group will consist of Senior Personnel from the three Project sites, the Project Collaborations (CMS & CDF), and Northwestern University. The Chair of Technical Committee will be Matt Crawford, who will also be a liaison member of the Steering Committee. The Technical Group will hold weekly meetings to cover the progress and status of the Project. The Steering Group will meet at least quarterly, and be responsible for producing quarterly reports. The Steering Group will also set Project milestone schedules.

5.6 Program of Work

We will present the status and results of our research into the experiments through the collaborations and in other appropriate forums. In this way we hope to benefit the common infrastructure services of the U.S. Physics Grid projects and more broadly influence other SciDAC projects.

We have established a general goal, or set of goals, for each year of the Project:

Year 1: We will perform the necessary work to move experiment data between our production storage systems via UltraNet, on a per-flow forwarding basis.

Year 2: We will move experiment data between Project sites at substantially higher data rates than via existing production network paths. We will make significant progress in automating the function of the LambdaStation.

Year 3: We will work to understand the overall place for dynamically provisioned optical network paths in the large scale data movement of our experiments. The functions of LambdaStation will be highly automated.

Within each of those one-year goals, we establish a set of milestones that we will meet in order to achieve our overall Project objective:

Year 1:

Ultranet Integration	Establish connectivity to UltraNet at each Project site. Liaise with UltraNet on scheduling & provisioning interface
Local network integration	Develop prototype LambdaStation
Application integration:	Build awareness of allocated bandwidth into SRM
Performance study:	Focus on monitoring and arrival process models
Overall Results:	Perform loopback testing from FNAL to SOX & Sunnyvale Forward flows on Ultranet between all sites, as well as to CERN

Year 2:

Ultranet Integration	Enhance connectivity to UltraNet to maximize performance.
Local network integration	Automate LambdaStation local network flow forwarding control. Automate inter-LambdaStation coordination. Automate LambdaStation UltraNet scheduling & provisioning. Introduce VO sensitivity to LambdaStation.
Application integration:	Adapt storage systems to schedule transfers according to alternate path availability.
Performance study:	Time series analysis, with initial work on hidden Markov models.
Overall Results:	Dynamic Ultranet path should significantly outperform existing production network path. Achieve behavior described in case study.

Year 3:

Ultranet Integration	Full integration to Ultranet.
Local network integration	Harden LambdaStation to production quality. Migrate what we've done to appropriate standards. Complete LambdaStation automation.
Application integration:	Consider where design has taken us, and what direction to follow.
Performance study:	Focus on hidden Markov models and locally weighted memory-based learning. Prototype per-flow automated forwarding, if practicable
Overall Results:	Lambda Station usable by experiments.

6 References

[AAA1]. Vollbrecht, et. Al., AAA authorization Framework, RFC2904, August 2000

[AAA2] . http://pcstats.cern.ch/DataTAG/UvA-AAA-Flyer.doc

[AAAQoS] L. Gommans, C. de Laat, B. van Oudenaarde and A. Taal, "Authorization of a QoS Path Based on Generic AAA", Technical Report University of Amsterdam, October 2002. *Future Generation Computer Systems*. [pdf]

[CDF] http:// www-cdf.fnal.gov/upgrades/computing/dh/chep03/ CdfRun2DataHandlingDesignTalk.pdf

[Chu1] Y. Chu, S. Rao, S. Seshan and H. Zhang, *A Case for End System Multicast***,** IEEE Journal on Selected Areas in Communication. Volume 20, Number 8.

[GMPLS1] R. Carlton, et.al, A framework for multiprotocol label switching, 1999. Draft-ietf-mpls-framework-03.txt

[GMPLS2] E. Rosen, A. Viswannthan, R. Callon, Mulitprotocol label switching architecture, IETF RFC-3031, January, 2001

[GMPLS3] K.Kompella, et. Al., OSPF extensions in support of generalized MPLA\S, draft-ietf-ccamp-gmpls-routing-00.txt

[dCache] Fermilab mass storage including dcache http://wwwhppc.fnal.gov/mss/mass_storage.html,see http://www.dcache.org/manuals/chep2003.pdf

[Dinda1] P. Dinda, *Online Prediction of the Running Time of Tasks*, Cluster Computing, Volume 5, Number 3, 2002.

[Dinda2] P. Dinda, D. O'Hallaron, *Host Load Prediction Using Linear Models* , Cluster Computing, Volume 3, Number 4, 2000.

[Dinda3] P. Dinda, B. Garcia, K. Leung, *The Measured Network Traffic of Compiler-Parallelized Programs*, Proceedings of the 30th International Conference on Parallel Processing (ICPP 2001).

[Emerge] http://www.icair.org/main_projects_infrast.html

[ESNet]http://www.es.net/hypertext/welcome/roadmap/Index.html

[FNALmeet http://www-isd.fnal.gov/wawg/2003-11-06-Ultranet-Meeting/

[Gara1] http://www-fp.mcs.anl.gov/qos/

[Globus] http://www.globus.org/

[GGF1] Optical Network Infrastructure for Grid https://forge.gridforum.org/projects/ghpn-rg/document/draft-ggf-ghpn-opticalnets-0/en/1

[GridDT] Ravot, S. Grid Data Transport, presented at the First International Workshop on Protocols for Fast Long-Distance Networks (CERN, Geneva, Feb. 3–4, 2003); see http://datatag.web.cern.ch/datatag/pfldnet2003/slides/ravot.ppt

[GridFTPV2] http://www-isd.fnal.gov/gridftp-wg/xmode/XModeProposal.pdf

[Guo1] L. Guo, and I. Matta, *The War Between Mice and Elephants,* ICNP 2001.

[Harchol1] M, Harchol-Balter, and A. Downey*, Exploiting Process Lifetime Distributions for Load Balancing,* SIGMETRICS 1996.

[ICAIR]http://www.icair.org/

[Kapadia1] N. Kapadia, J. Forter, and C. Brodley. *Predictive Application-Performance Modeling in a Computational Grid Environment*, HPDC 1999.

[Kroeger1] T. Kroeger, and D. Long, *Predicting File-system Actions from Prior Events*, USENIX 1996.

[LCG] http://www.griphyn.org/news/

[NetL1] S. Low et al., Caltech's Network Laboratory, see http://netlab.caltech.edu and Jin, C., Wei, D., Low, S., Buhrmaster, G., Bunn, J., Choe, D., Cottrell, R., Doyle, J., Newman, H., Paganini, F., Ravot, S., and Singh, S. FAST kernel. Background theory and experimental results. Presented at the First International Workshop on Protocols for Fast Long-Distance Networks (CERN, Geneva, Switz., Feb. 3–4, 2003); http://netlab.caltech.edu/pub/papers/pfldnet.pdf

[MonA1] H.B. Newman, I.C.Legrand, P. Galvez, R. Voicu, C. Cirstoiu, MonALISA Distributed Monitoring Service Architecture, http://arxiv.org/ftp/cs/papers/0306/0306096.pdf, March 2003.

[Myers1] A. Myers, P. Dinda, H. Zhang, *Performance Characteristics of Mirror Servers on the Internet*, Proceedings of Infocom 1999, New York City, New York, March 1999, pp. 304-312

[Photonic] Mambretti et al., "The Photonic TeraStream: enabling next generation applications through intelligent optial networking at iGRID2002", Furute Generation Computer Systems 19 (2003) 897-908, Elsivier.

[Qiao1] Y. Qiao, J. Skicewicz, and P. Dinda, *Multiscale Predictability of Network Traffic*, Technical Report NWU-CS-02-13, Department of Computer Science, Northwestern University, October, 2002.

[SRM] Storage Resource Manager Project http://sdm.lbl.gov/indexproj.php?ProjectID=SRM, see also http:// sdm.lbl.gov/srm-wg/doc/SRM.spec.v2.1.final.pdf

[Sundararaj1] A. Sundararaj, and P. Dinda, *Towards Virtual Networks for Virtual Machine Grid Computing*, USENIX, 2004.

[Ultralight] http://ultralight.caltech.edu/

[Ultranet] http://www.csm.ornl.gov/ultranet/

[Wolski1] R. Wolski, *Dynamically Forecasting Network Performance using the Network Weather Service*, Cluster Computing Volume 1, pp. 119-132, January, 1998.

[Willinger1] W. Willinger, M. Taqque, R. Sherman, D. Wilson, *Self-similarity Through High-variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level*, IEEE/ACM Transaction on Networking, Volume 5, Number 1, 1997.

[Zhang1] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker, *On the Characteristics and Origins of Internet Flow Rates,* SIGCOMM 2002.